Towards Computational Historiographical Modeling

Michael Piotrowski
University of Lausanne, Switzerland
michael.piotrowski@unil.ch @mxp@mastodon.acm.org

2024-09-12

Abstract

Digital corpora play an important, if not defining, role in digital history and may be considered as one of the most obvious differences to traditional history. Corpora are essential for the use of computational methods and thus for the construction of computational historical models. But beyond their technical necessity and their practical advantages, their epistemological impact is significant. While the traditional pre-digital corpus is often more of a potentiality, a mere "intellectual object," the objective of computational processing requires the corpus to be made explicit and thus turns it into a "material object." Far from being naturally given, corpora are constructed as models of a historical phenomenon and therefore have all the properties of models. Moreover, following Gaston Bachelard, I would argue that corpora actually construct the phenomenon they are supposed to represent; they should therefore be considered as phenomenotechnical devices.

1 Introduction

What do I mean by "computational historiographical modeling"? This becomes clearer when you know my definition of computational humanities.

Slide 2

~~Digital~~ Computational Humanities

1. The **applied computational humanities** are concerned with the construction of formal—computational—models of the phenomena studied by their "mother disciplines," as well as the methodology of their construction.
2. The **theoretical computational humanities** study the general properties of formal—computational—models in the humanities at a higher level of abstraction.

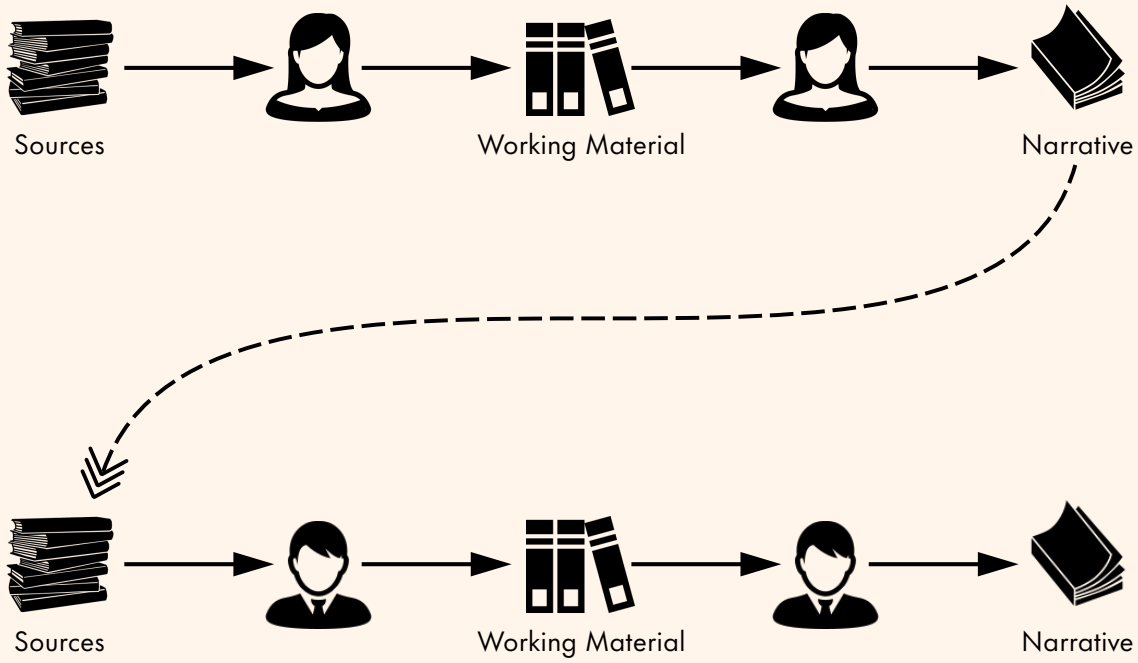
2 Computational Models in the Humanities Research Process

All research creates models; as Granger ([1960] 1967) notes, the goal of any science (natural or other) is to build *coherent and effective models of the phenomena they study*. However, in the traditional historical research process, the models are rarely explicit, let alone formal. It works somewhat like this:

- Scholar reads and interprets primary and secondary **sources**.
- Facts and insights are recorded as **working materials** in a variety of forms (on paper or electronically, as text, in spreadsheets, databases, etc.).
- Using the working materials, scholar constructs **mental model** to answer research question and describes the model in a **narrative**.

Slide 3

Traditional Research Process

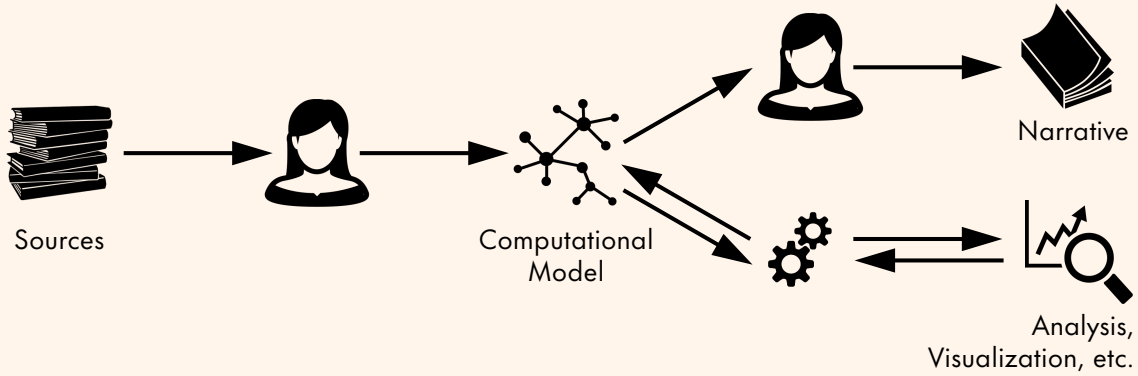


If you want to build on previous research, you essentially have to interpret the narrative and reconstruct the underlying model.

One potential way of imagining a computational research process is to imagine a computational model in the place of the working material: while consulting the sources, you build an explicit model that formally describes how you think that things relate to each other.

Slide 4

Potential Computational Research Process

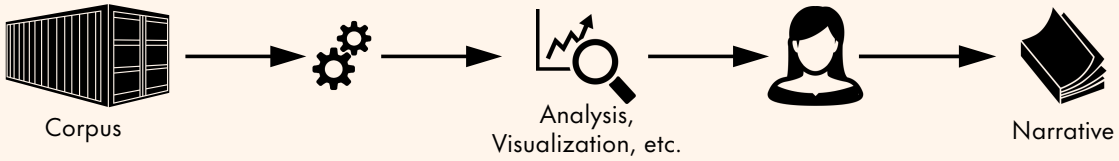


One additional advantage would be that you could directly compare or combine different such models.

Now, in actual practice, the research process in ~~digital~~ computational history typically looks more like this:

Slide 5

Typical ~~Digital~~ Computational Research Process



When we look for epistemological differences between "traditional" and digital history, the *corpus*—stands out. Of course, historians have always created and studied collections of traces, in particular documents, but sometimes also other artifacts, and have built their narratives on the basis of these collections. This is a significant aspect of scholarship and in some sense constitutes the difference between historical and literary narratives: historical narratives are supposed to be grounded (in some way) in the historical facts represented by the respective corpus.

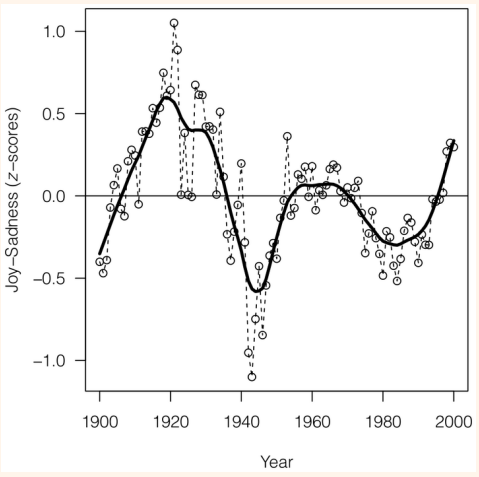
Nevertheless, the relation between such a corpus and the narrative is traditionally rather unclear. Not only is the corpus necessarily incomplete (and uncertain), but it's typically only "virtual." As Mayaffre (2006, 20) puts it, in the humanities corpora traditionally tend to be potentialities rather than realities: one *could* go and consult a certain document in some archive, but this may only be rarely done, and the corpus may thus have never been anything but an "intellectual object."

Machine-readable digital corpora—that is, what we mean by corpora today—have brought about major changes. Most of the time, it is their practical advantages that are highlighted: they are easier to store, they are (at least potentially) accessible from anywhere at any time, and they can be processed automatically. This, in turn, enables us to apply new types of analysis and thus to ask and study new research questions. What tends to be overlooked, though, is the epistemological impact of machine-readable corpora in history. The notion of corpus in digital history (and in digital humanities in general) is heavily influenced by the notion of corpus in computational linguistics: a large but finite collection of digital texts. Mayaffre (2006, 20) hints at the epistemological impact when he notes that, on the one hand, digitization dematerializes the *text* in that it is lifted from its previous support, but on the other hand, materializes the *corpus* more rigorously than before.

This is, of course, a precondition for more rigorous types of analysis, notably computational analyses, and—eventually—the construction of computational historical models. However, this raises a number of epistemological and methodological questions. In computational linguistics, a corpus is essentially considered a statistical sample of language. Historical corpora typically differ from linguistic corpora, both in its relation to the research objects, the research questions, and to the expected research findings. They also differ in the way they are constructed.

Slide 6

According to Acerbi et al. (2013), moods have followed broad historical trends, including a "sad" peak corresponding to World War II, and two "happy" peaks, one in the 1920s and one in the 1960s. There is also a "sad" period starting in the 1970s, with an increase in "happiness" in the later years of the data set.



Given the central role of corpora in digital history, I think we need to study them and the roles they play in order to avoid the production of research that is formally rigorous but historically meaningless (or even nonsensical).

3 Corpora as Models

As Granger ([1960] 1967) notes, the goal of any science (natural or other) is to build *coherent and effective models of the phenomena they study*.

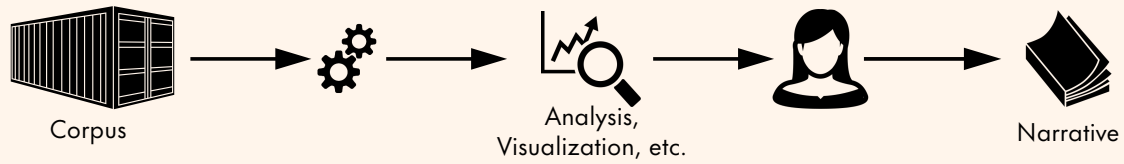
Slide 7

> Quant à l'intentionnalité scientifique, à la visée, nous l'avons déjà définie comme construction de *modèles cohérents et efficaces du phénomène*.
> —Granger ([1960] 1967, 215)

So, if we look at this pipeline, where's the model?

Slide 8

Typical ~~Digital~~ Computational Research Process



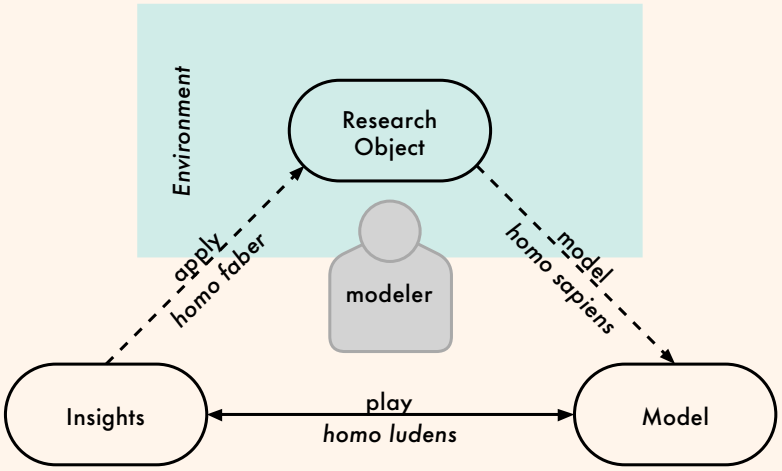
Of course, there's not just *one* model. But people tend to focus on the right half and ignore what might be the most important one: *the corpus*.

Thus, and as I have argued before (Piotrowski 2019), a corpus should be considered a model in the sense of Leo Apostel, who asserted that "*any subject using a system A that is neither directly nor indirectly interacting with a system B to obtain information about the system B*, is using A as a model for B" (Apostel 1961, 36, emphasis in original). Creating a corpus thus means constructing a model, and modelers consequently have to answer questions such as: What is it that I am trying to model? In what respects is the model a reduction of it? And for whom and for what purpose am I creating the model?

These are not new questions: every time historians select sources, they construct models, even before any detailed analysis. However, machine-readable corpora are not only potentially much larger than any material collection of sources—which is already not inconsequential—but also have important epistemological consequences. The larger and the more "complete" a corpus is, the greater the danger to succumb to an "implicit essentialism" (Mothon 2010, 19) and to mistake the model for the original, a fallacy that can frequently be observed in the field of cultoromics (Michel et al. 2011), when arguments are being made on the basis of the Google Books Ngram Corpus.

The same then goes for any analysis of a corpus: if the corpus is "true," so must be the results of the analysis; if there is no evidence of something in the corpus, it did not exist. This allure is even greater when the analysis is done automatically and in particular using opaque quantitative methods: as the computational analysis is assumed to be completely objective, there seems to be no reason to question the results—they merely need to be interpreted, which leads us to some kind of "digital positivism." To rephrase Fustel de Coulanges (Monod 1889, 278), "Ne m'applaudissez pas, ce n'est pas moi qui vous parle ; ce sont les données qui parlent par mes courbes."

Slide 9



An analysis of a corpus will *always* yield results; the crucial question is whether these can tell us anything about the phenomenon it aims to model. So, my point is that corpora are not naturally occurring but intentionally constructed. A corpus is *already* a model and thus not epistemologically neutral.

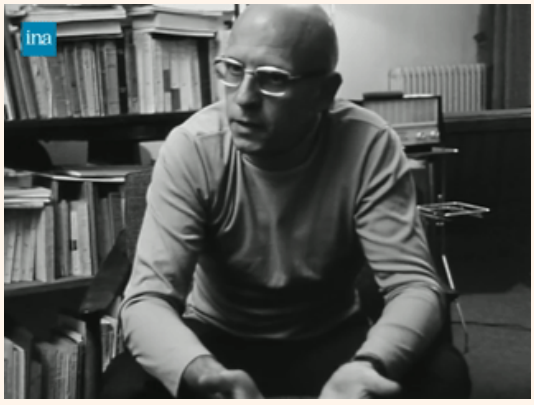
4 Corpora as Phenomenotechnical Devices

This may not sound particularly exciting: sure, corpora are constructed; yes, they can be biased, etc. However, there's more to it. Corpora, in particular as they are used in digital history, are not just passive reflections of the historical phenomena under study. In fact, they *create* these phenomena.

A good way to describe this is Gaston Bachelard's notion of *phenomenotechnique* (Bachelard [1934] 1968). Bachelard originally developed this notion, which treats scientific instruments as "materialized theories," as a way to study the epistemology of modern physics, which goes far beyond what is directly observable. The humanities also and even primarily deal with phenomena that are not directly observable, but only through artifacts, in particular texts. They thus have also always constructed the objects of their studies through, for example, the categorization and selection of sources and the hermeneutic postulation and affirmation of phenomena.

Slide 10

> La véritable phénoménologie scientifique est donc bien essentiellement une phénoménotechnique. […] Elle s'instruit par ce qu'elle construit.
> —Bachelard ([1934] 2020, 35)



However, only the praxis has been codified to some extent as "best practices," such as source criticism. What history (and the humanities in general) traditionally do not have is something that corresponds to the scientific instrument.

This changes with digitalization and datafication: phenomena are now constructed and modeled through data and code, and (like in the sciences), the computational model takes on the role of the instrument and "sits in the center of the epistemic ensemble" (Rheinberger 2005, 320). Corpora

are then, methodologically speaking, phenomenotechnical devices and form the basis and influence how we build, understand, and research higher-level concepts—which at the same time underly the construction of the corpus. In short: a corpus produces the phenomenon to be studied.

5 What Kind of Models are Historical Corpora?

Slide 11

What Kind of Models are Historical Corpora?

I've said that historical corpora are not like linguistic corpora, which are essentially statistical models. This raises the question, what kind of models are historical corpora?

In the literature on modeling we find many classifications of models (e.g., Штофф 1966; Stachowiak 1973; Varenne 2017). I don't find most of these attempts particularly useful. The problem is that they generally treat models as entities, as artifacts, or objects, and then search for common properties, for some kind of essence. But this is difficult, because of the huge diversity of models, which then leads some researchers to the conclusion that the notion of "model" is useless and should be abandoned (Veit 2023).

However, models are better understood as relations, so if we want to classify models, it is more useful to classify them with respect to their *relation to the original* that they represent.

Thus we can say that statistical models are characterized by the relationship between a population and a sample of this population. This is a specific and formally defined relationship, and forms the basis of statistics.

Slide 12

> A map is *not* the territory it represents, but, if correct, it has a *similar structure* to the territory, which accounts for its usefulness. —Korzybski (1933, 58)

A map, in contrast, as Korzybski (1933, 58) famously remarked, is useful because it has a similar structure to the territory—a different relation.

Now, a collection of historical documents that we take as evidence for some historical phenomenon is neither a statistical sample, nor does it have structural similarity to the phenomenon. So what is the relationship here?

One approach could be to say that what we have here is a trace (this is again quite similar to what we have in law, but also in the natural sciences): The relationship between a trace and the phenomenon that is supposed to have created that phenomenon is a causal relationship.

6 Conclusion

I have tried to outline some of the background and the motivation for the project *Towards Computational Historiographical Modeling: Corpora and Concepts*, which is part of a larger research program.

So far, digital history (and digital humanities more generally) has largely contented itself with borrowing methods from other fields and has developed little methodology of its own. The focus on "methods and tools" represents a major obstacle towards the construction of computational models that could help us to obtain new insights into *humanities* research questions rather than just automate primarily quantitative processing—which is, without doubt, useful, but inherently limited, given that the research questions are ultimately qualitative.

Regardless of the application domain, digital humanities research tends to rely heavily on *corpora*, i.e., curated collections of texts, images, music, or other types of data. However, both the epistemological foundations—the underlying concepts—and the epistemological implications have so far been largely

ignored. I have proposed to consider corpora as *phenomenotechnical devices* (Bachelard [1934] 1968), like scientific instruments: corpora are, on the one hand, models of the phenomenon under study; on the other hand, the phenomenon is *constructed* through the corpus.

We therefore need to study corpora as models to answer questions such as: How do corpora model and produce phenomena? What are commonalities and differences between different types of corpora? How can corpora-as-models be formally described in order to take their properties into account for research that makes use of them?

The overall goal of the project is to contribute to theory formation in digital history and digital humanities, and to help us move from project-specific, often ad hoc, solutions to particular problems to a more general understanding of the issues at stake.

Acknowledgements

This research was supported by the Swiss National Science Foundation (SNSF) under grant no. 105211_204305.

References

Acerbi, Alberto, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. "The Expression of Emotions in 20th Century Books." *PLoS ONE* 8 (3): e59030. https://doi.org/10.1371/journal.pone.0059030.

Apostel, Leo. 1961. "Towards the Formal Study of Models in the Non-Formal Sciences." In *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences*, edited by Hans Freudenthal, 1–37. Dordrecht: Reidel. https://doi.org/10.1007/978-94-010-3667-2_1.

Bachelard, Gaston. (1934) 1968. *Le nouvel esprit scientifique*. 10th ed. Paris: Les Presses universitaires de France.

———. (1934) 2020. *Le nouvel esprit scientifique*. Edited by Vincent Bontems. 1re édition critique. Paris: Les Presses universitaires de France.

Granger, Gilles-Gaston. (1960) 1967. *Pensée formelle et sciences de l'homme*. Nouvelle éd. augmentée d'une préface. Paris: Aubier-Montaigne.

Korzybski, Alfred. 1933. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. Lancaster, PA: International Non-Aristotelian Library Publishing Company. https://n2t.net/ark:/13960/t6c261n93.

Mayaffre, Damon. 2006. "Philologie et/ou herméneutique numérique: nouveaux concepts pour de nouvelles pratiques?" In *Corpus en lettres et sciences sociales: des documents numériques à l'interprétation. Actes du XXVIIe Colloque d'Albi "Langages et signification"*, edited by François Rastier and Michel Ballabriga, 15–25. CALS-CPST. https://hal.science/hal-00551477.

Michel, Jean-Baptiste, Yuan K. Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. https://doi.org/10.1126/science.1199644.

Monod, Gabriel. 1889. "M. Fustel de Coulanges." *Revue historique* 42 (2): 277–85. https://www.jstor.org/stable/40938008.

Mothon, Bernard. 2010. *Modélisation et vérité*. Paris: Archétype82.

Piotrowski, Michael. 2019. "Historical Models and Serial Sources." *Journal of European Periodical Studies* 4 (1): 8–18. https://doi.org/10.21825/jeps.v4i1.10226.

Rheinberger, Hans-Jörg. 2005. "Gaston Bachelard and the Notion of 'Phenomenotechnique'." *Perspectives on Science* 13 (3): 313–28. https://doi.org/10.1162/1063614054774288026.

Stachowiak, Herbert. 1973. *Allgemeine Modelltheorie*. Wien, New York: Springer.

Varenne, Franck. 2017. *Théories et modèles en sciences humaines: Le cas de la géographie*. Paris: Éditions Matériologiques.

Veit, Walter. 2023. "Model Anarchism." *THEORIA. An International Journal for Theory, History and Foundations of Science* 38 (2): 225–45. https://doi.org/10.1387/theoria.23849.

Штофф, Виктор Александрович. 1966. *Моделирвание и Философия*. Москва: Наука.